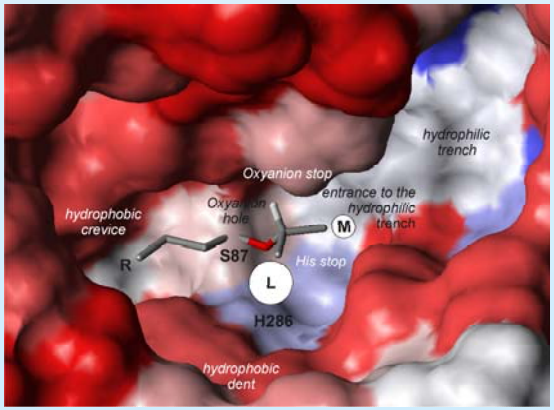
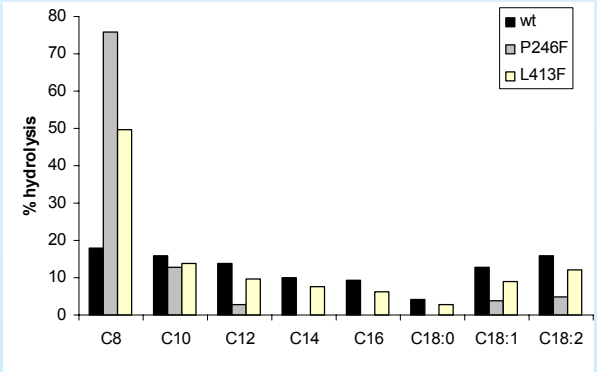
Datenbanken in der Bioinformatik

Jürgen Pleiss

Institut für Technische Biochemie, Universität Stuttgart


```
                                MSO
Asp.oryzae          LKPEHSDYKIVVVGHSLGAAIASLAAADLRT
Pen.camembertii     VVAQNPNYELVVVGHSLGAAVATLAATDLRG
Rh.oryzae           QLTAHPTYKVIVTGHSLGGAQALLAGMDLYQ
1LGY-A -B -C        HHHHHHH EEEEEEE HHHHHHHHHHHHHHH
Rh.miehei           QFKQYPSYKVAVTGHSLGGATALLCALDLYQ
closed (3TGL)       HHHHHHH EEEEEEE HHHHHHHHHHHHHHH
open (4TGL)         HHHHHHH EEEEEEE HHHHHHHHHHHHHHH
Hum.lanuginosa      AVREHPDYRVVFTGHSLGGALATVAGADLRG
closed (1TIB)       HHHHHHH EEEEEEE HHHHHHHHHHHHH
Fus.heterosporum    AKTANPTFKFVVTGHSLGGAVATIAAAYLRK
                          . :.. ..*****.* * :..  *
```





Datenaustausch vor 1990 ...

lokale Zentren (Bibliotheken, Rechenzentren), die lokal Daten (Publikationen, Datenbanken) zur Verfügung stellen

Kommunikation per Post:
Verschicken von Zeitschriften, Büchern, Bändern

Internet : Netzwerk von Computern mit gemeinsamem Kommunikationsprotokoll
Services: e-Mail, FTP, remote login

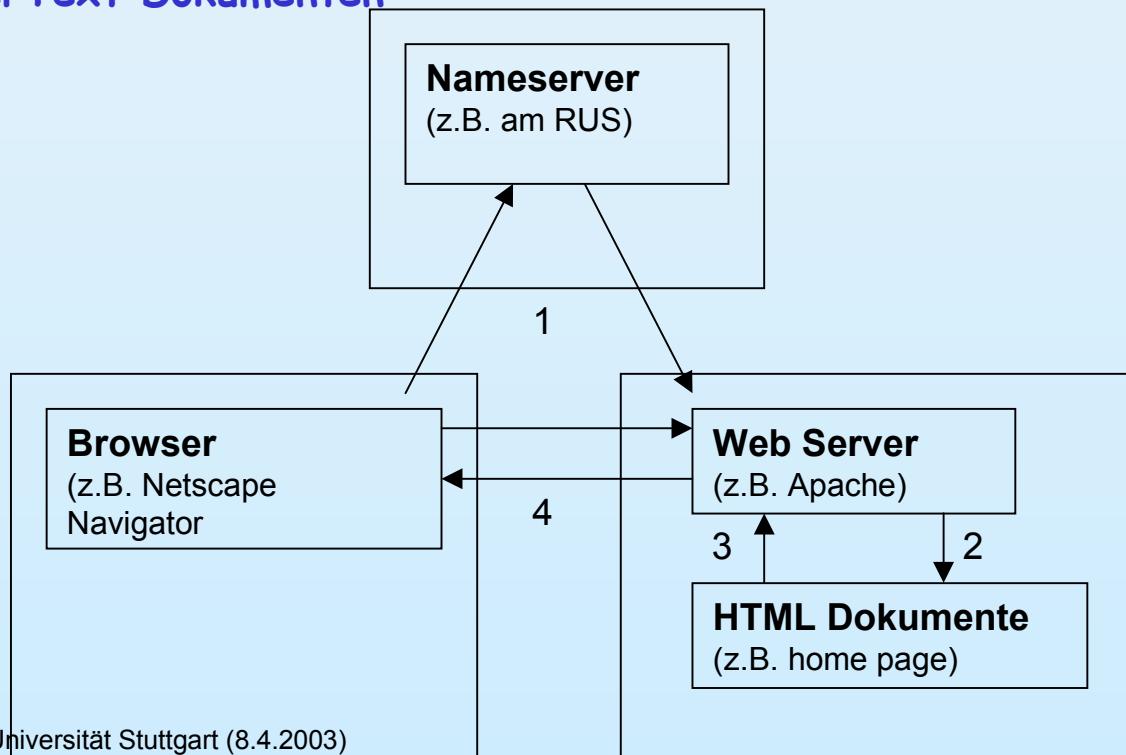
Seit 1989: World Wide Web

1989 Tim Berners-Lee (CERN, Genf): World Wide Web

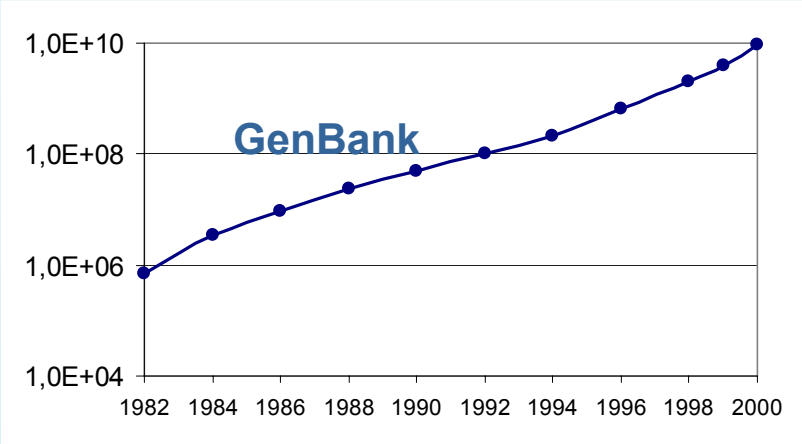
Hyper Text Markup Language (HTML)

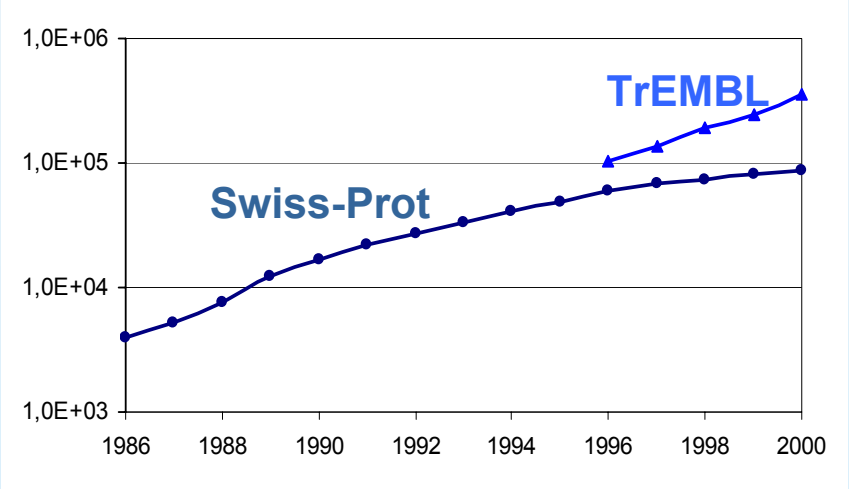
Netzwerk von verbundenen HyperText Dokumenten

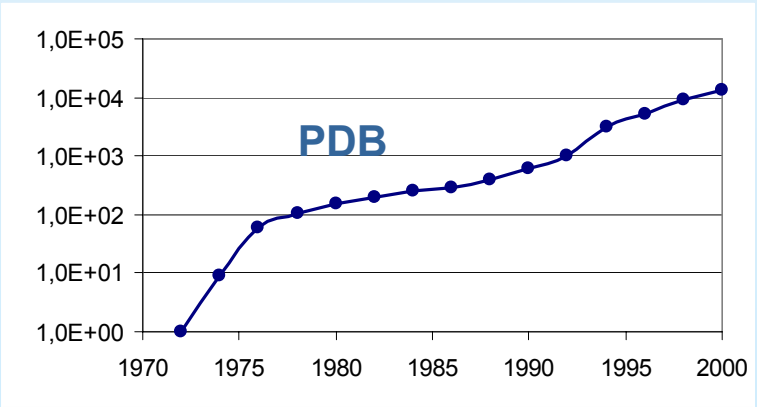
erster textbasierter Browser



Austausch biologischer Daten



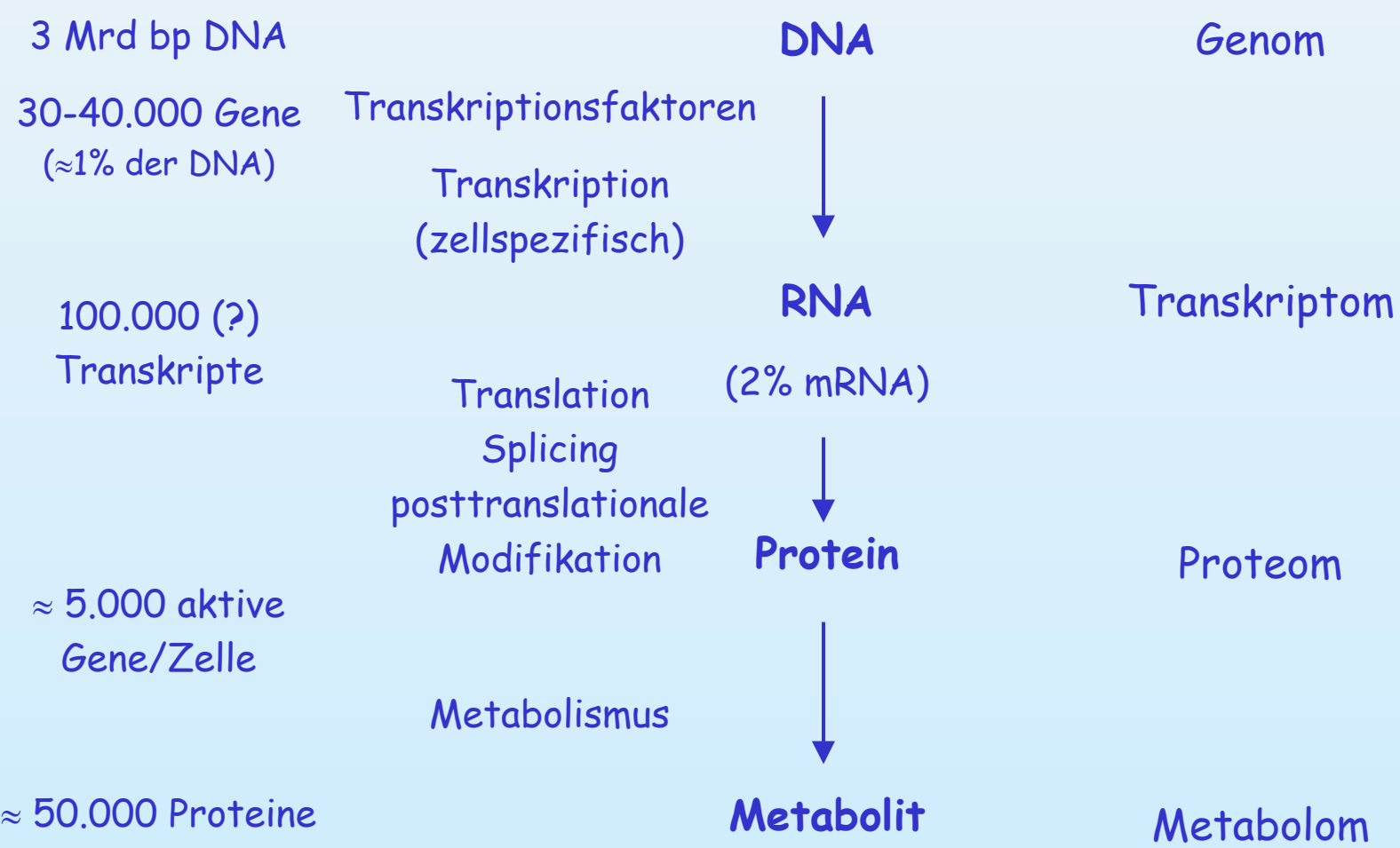




Oktober 2002

20 Mrd Basen	**+ 60%** pro Jahr
680.000 TrEMBL Einträge	**+ 40%** pro Jahr
115.000 Swiss-Prot-Einträge	**+ 20%** pro Jahr
19.000 PDB-Einträge	**+ 25%** pro Jahr

Biologische Daten: Genomik und Post-Genomik



Was macht Bioinformatik?

Regulation:

- zeitlich
- räumlich
- zellspezifisch
- externe Signale
- Entwicklung

-omics : genomics, transcriptomics, proteomics, metabolomics

Wie funktioniert / entwickelt sich / reagiert eine Zelle / ein Organismus?

→ Funktionelle Genomik, Systembiologie

Ein Beispiel: die *Lipase Engineering Database*

enthält

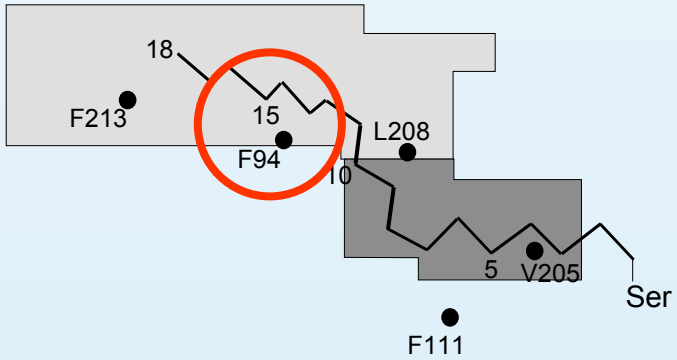
- Proteinsequenz
- Struktur
- Informationen über einzelne Positionen
- Einteilung nach Familien
- Sequenzvergleich
- Phylogenetische Bäume

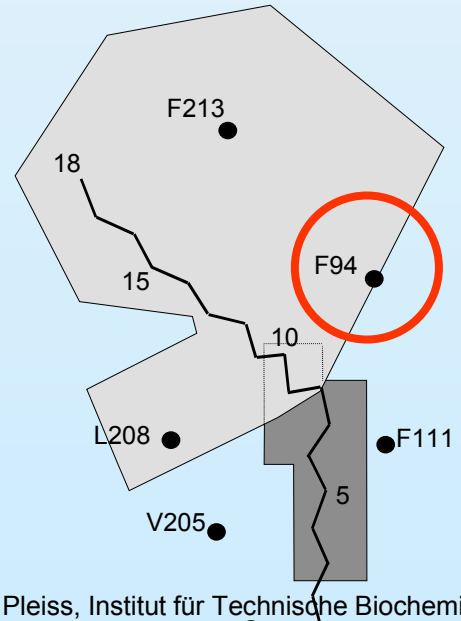
einer Enzymfamilie (>1200 Sequenzen)

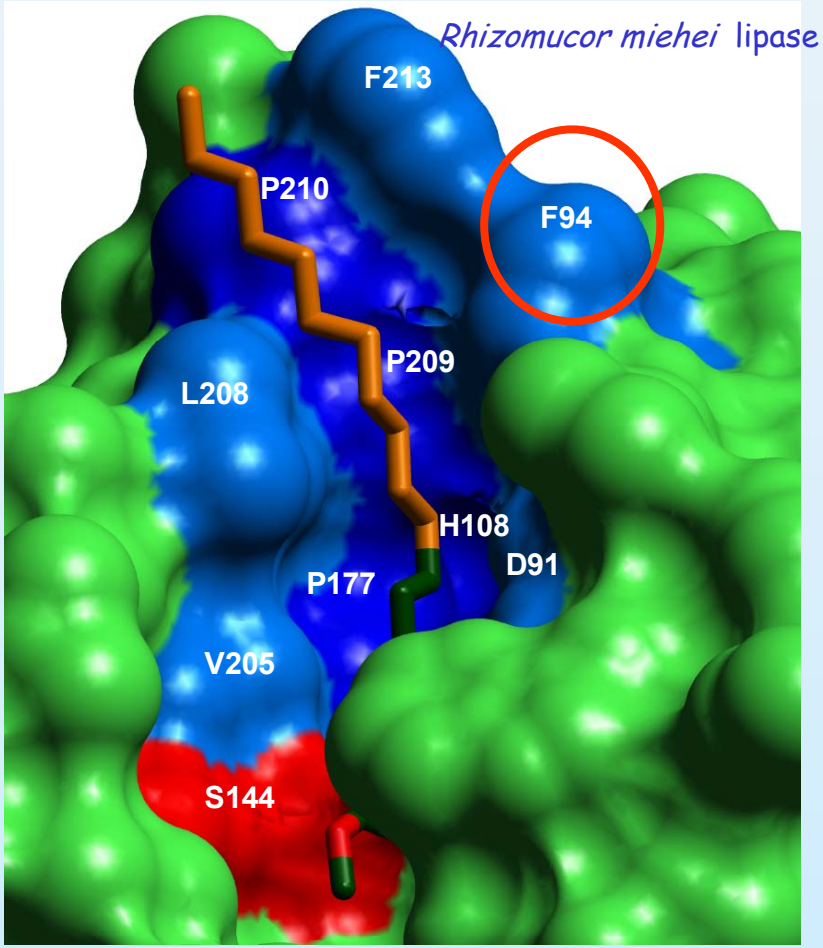
Fragen: Was haben diese Enzyme gemeinsam? Was unterscheidet sie?

Wie kann man die Funktion in der Sequenz lesen?

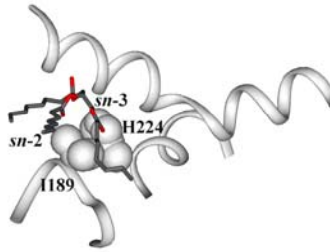
Sequenz → Struktur → Eigenschaften

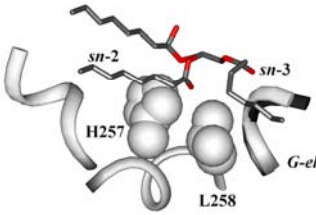


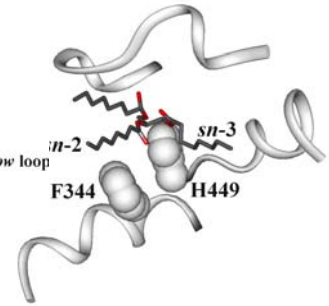




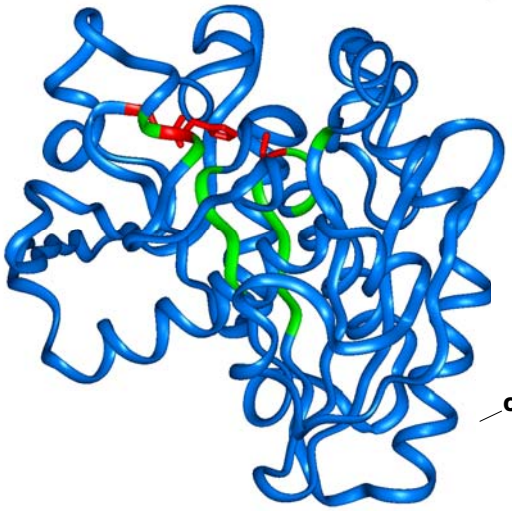
Ziel: Verständnis der biochemischen Eigenschaften

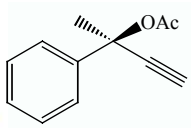


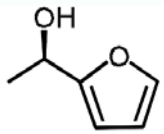


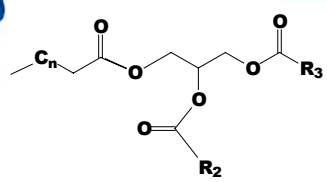


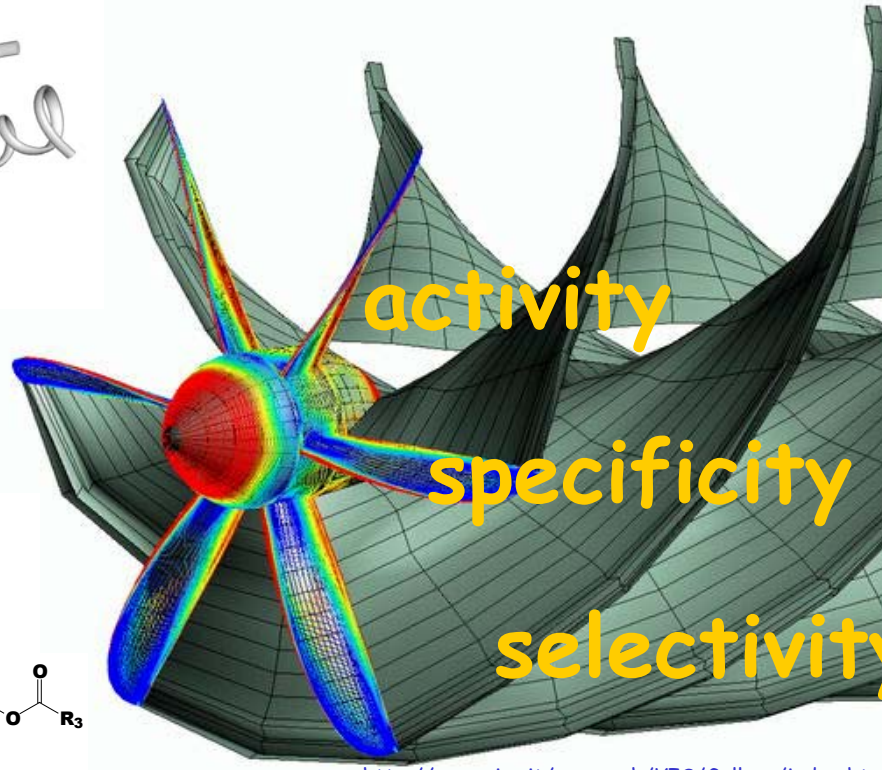

```
ITTFT-SLL
IKTWS-TLI
SYDFSDSTI
LYSFEDSGV
   :  : :

TNGYVLRSDK
TNAMVARGDS
TAGYIAVDHT
VTGFLALDNT
. . :  ...

TIYLVFRGTNSF
TIYIVFRGSSSI
AVVLAFRGSYSV
LIVLSFRGSRSI
 : : ***: *.
```







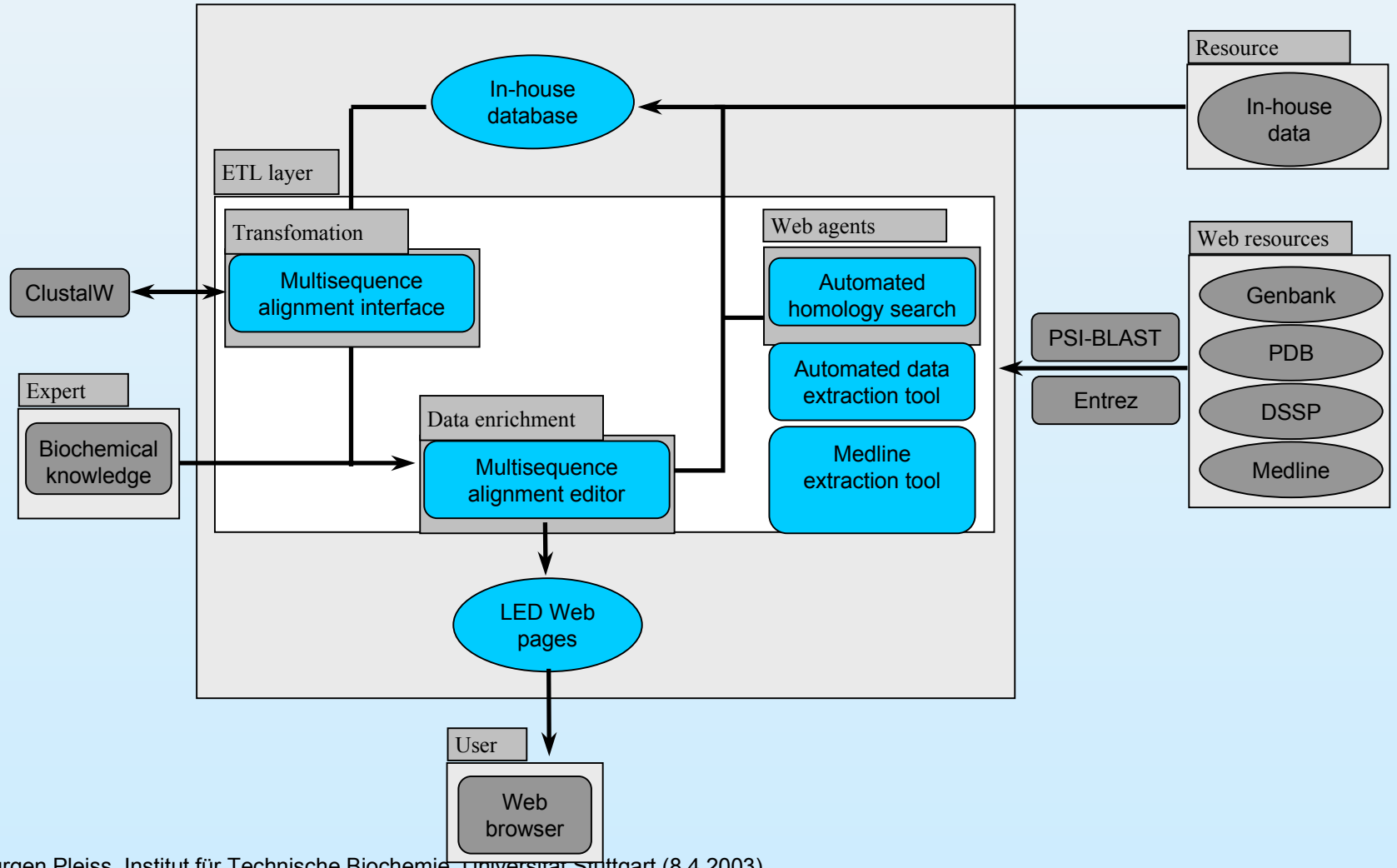




http://www.cira.it/research/VIS/Gallery/index.htm

```
GFLSSYEQVVNDYFPVVQEQLTAHPTYKVIVTGHSLGGAQALLAGMDLYQREPRAKLYAYASPRVGNAALAKYITAQ--GNNFRFTHTN-DPVPKLPLLSMGYVHVSPEYWITS
GFLDSYGEVQNELVATVLDQFKQYPSYKVAVTGHSLGGATALLCALDLYQREEGLSIFTVGGPRVGNPTFAYYVEST--GIPFQRTVHKRDIVPHVPPQSFGFLHPGVESWIKS
GFTSSWRSVADTLRQKVEDAVREHPDYRVVFTGHSLGGALATVAGADLR-----LFLYTQGQPRVGDPAFANYVVST--GIPYRRTVNERDIVPHLPPAAFGFLHAGEEYWITD
GFWSSWKLVRDDIIKELKEVVAQNPNYELVVVGHSLGAAVATLAATDLR-----IDVFSYGAPRVGNRAFAEFLTVQTGGTLYRITHTN-DIVPRLPPREFGYSHSSPEYWIKS
** .*:  * :     : : .   * *.: ..*****.* * :.. **         ::: . ***.:  :* ::    *  :: *  . * **::*   :*: * . * ::.
```

Lipase Engineering Database : Design



THE LIPASE ENGINEERING DATABASE

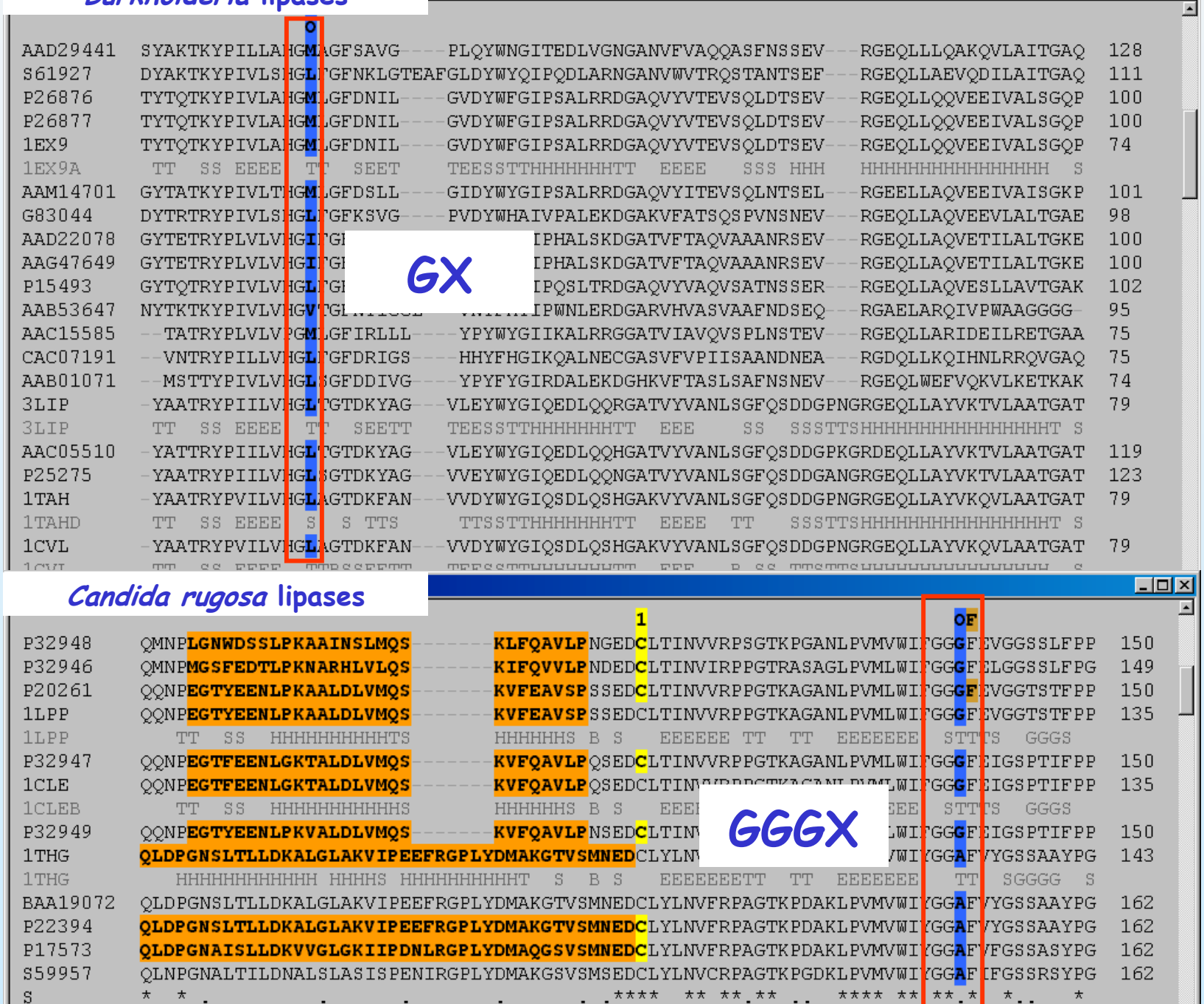
University of Stuttgart

www.led.uni-stuttgart.de

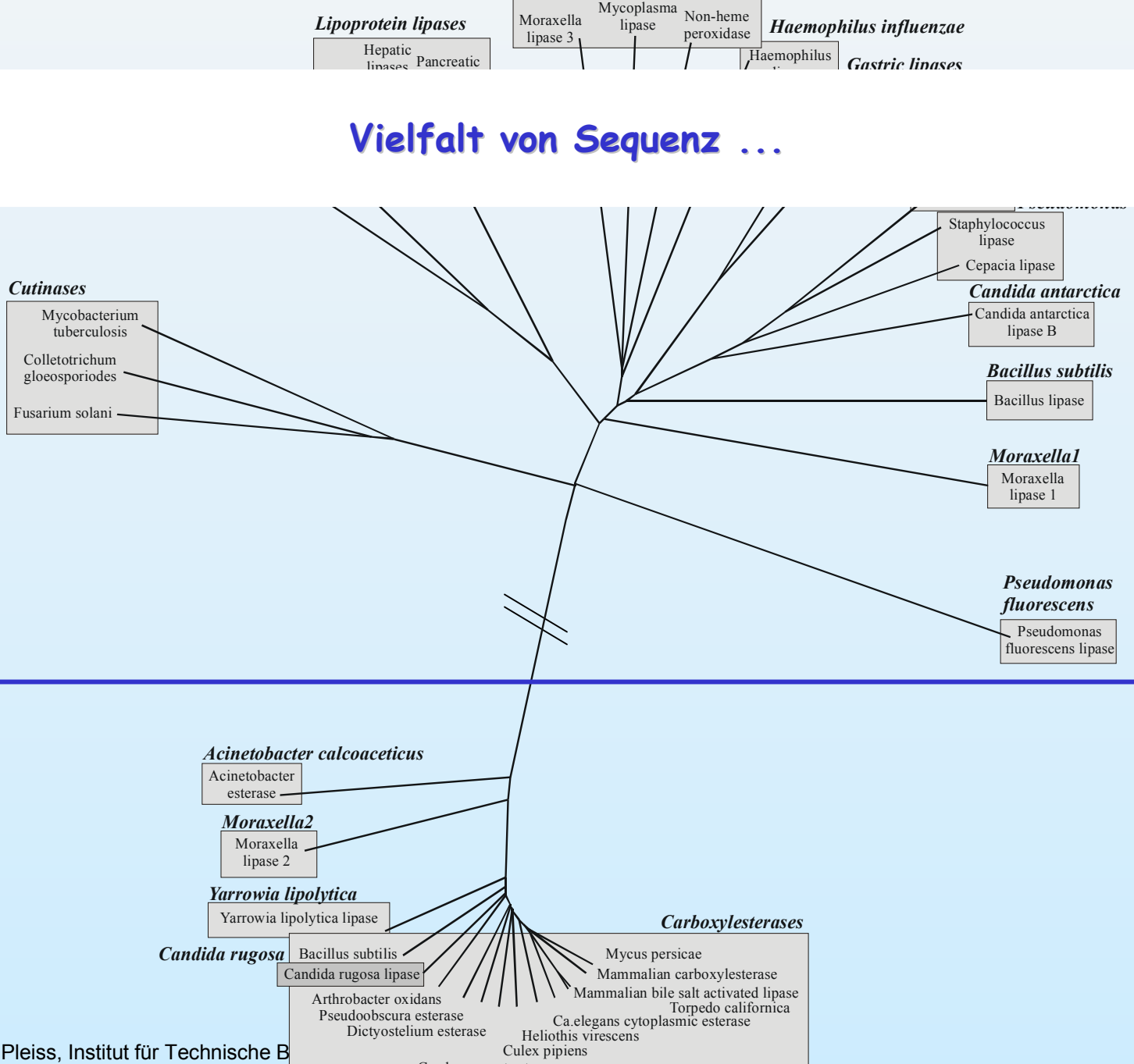
Browse classification for GX class

Superfamilies	11
Homologous families	22
Proteins	376
Sequences	600
Structures	124
Links	Structure table

Superfamily	Homologous family	Links
Bacillus lipases Proteins 6 Sequences 9 Structures 2	Bacillus lipases	Alignment Tree
Burkholderia lipases Proteins 41 Sequences 68 Structures 15 Alignment Tree	Burkholderia lipases	Alignment Tree
	Staphylococcus lipases	Alignment Tree
Candida antarctica lipase Proteins 1 Sequences 8 Structures 7	Candida antarctica lipase B	Alignment Tree
Cutinases Proteins 21 Sequences 71 Structures 44 Alignment Tree	Botryotinia cutinases	Alignment Tree
	Colletotrichum cutinases	Alignment Tree
	Fusarium cutinases	Alignment Tree
	Mycobacterium cutinases	Alignment Tree
Filamentous fungi lipases Proteins 36 Sequences 70 Structures 29	Rhizomucor lipases	Alignment Tree
	Saccharomyces lipases	Alignment Tree

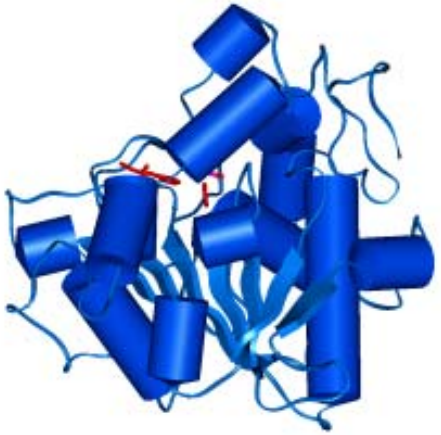

Burkholderia lipases
GX
Candida rugosa lipases
GGGX

Vielfalt von Sequenz ...

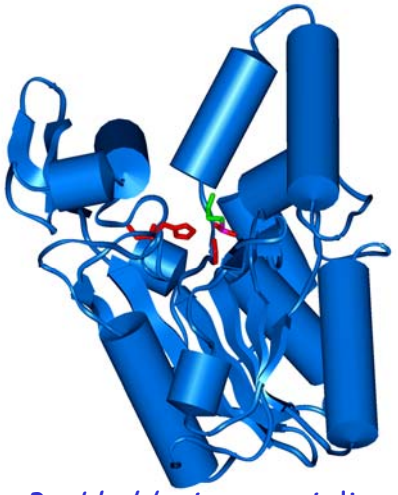


... und Struktur

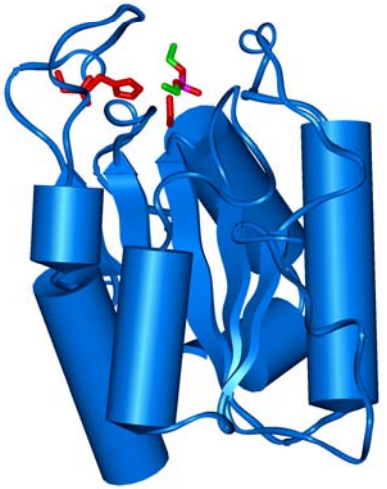
Rhizomucor miehei lipase



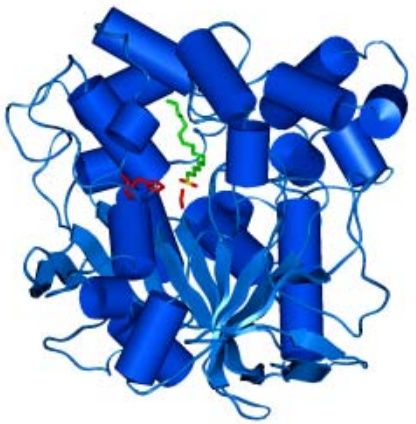
Candida antarctica lipase B



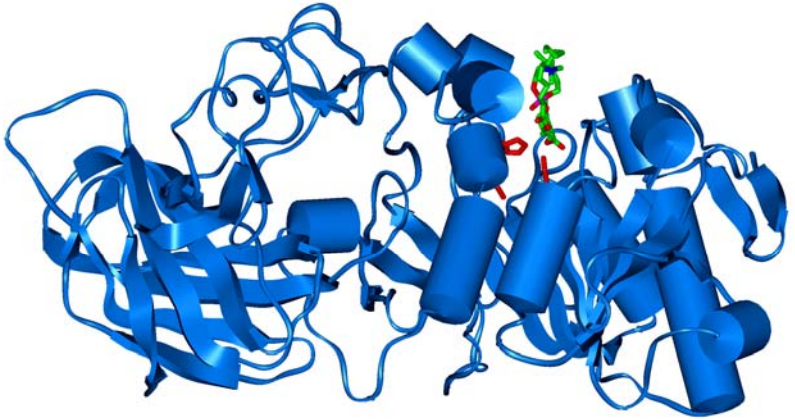
Burkholderia cepacia lipase



Fusarium solani cutinase

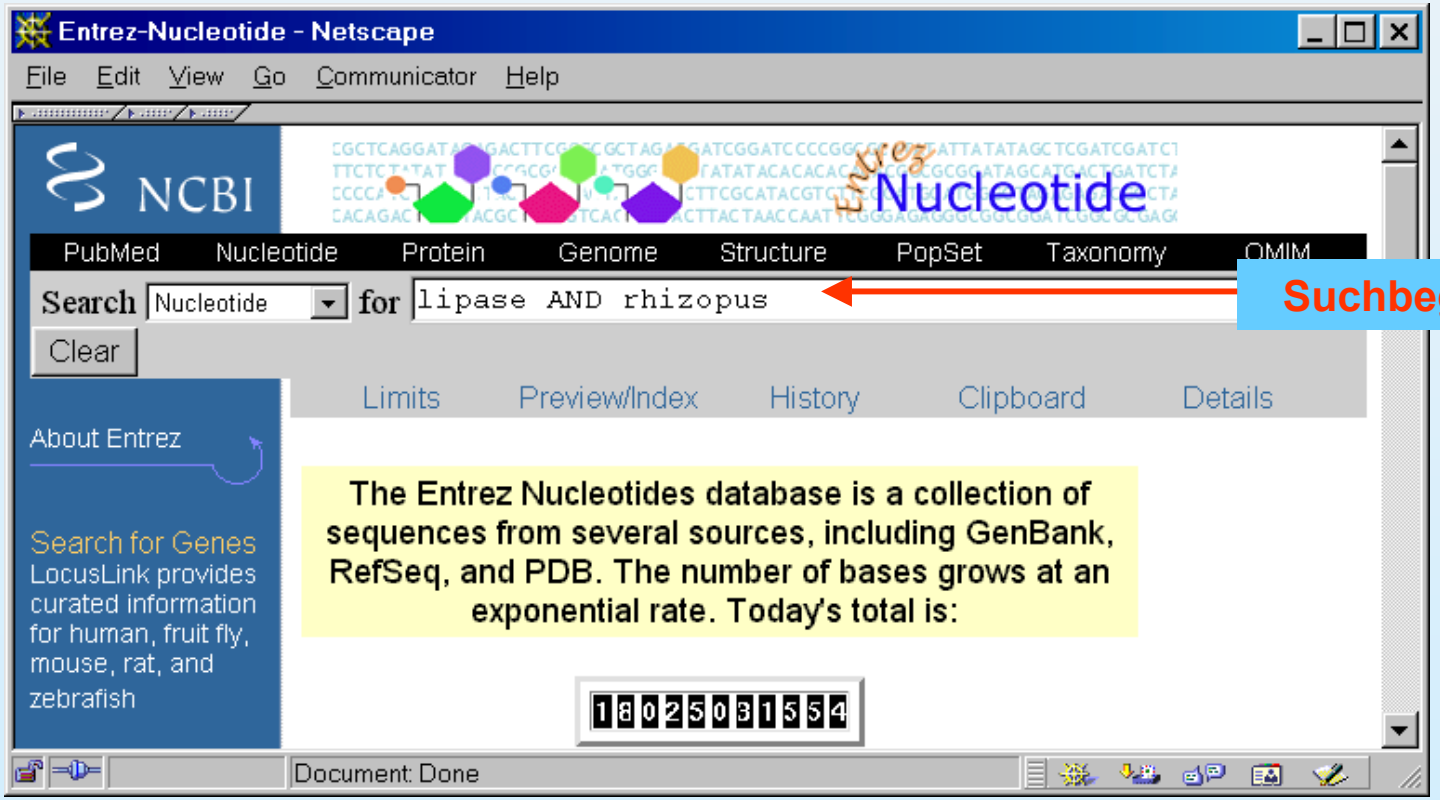


Thermomyces lanuginosa lipase



Human pancreas lipase

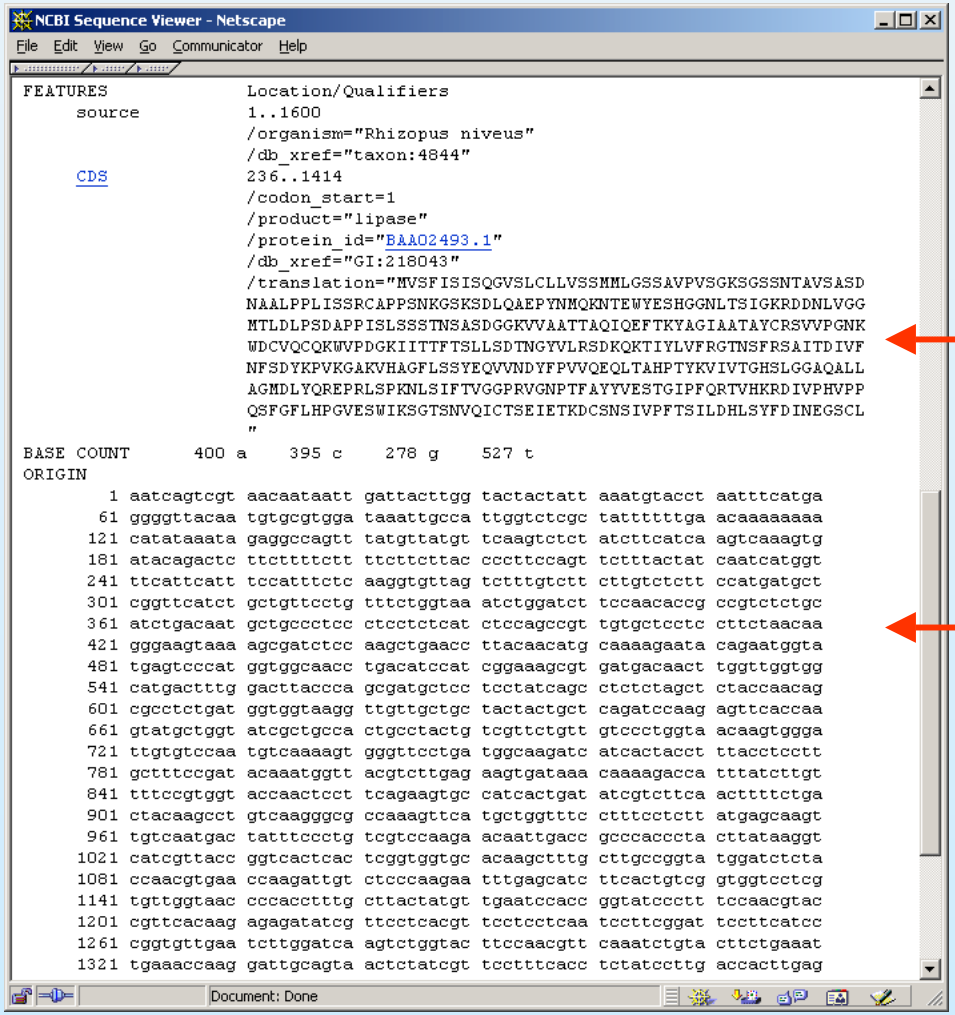
Sequenzdatenbank: GenBank



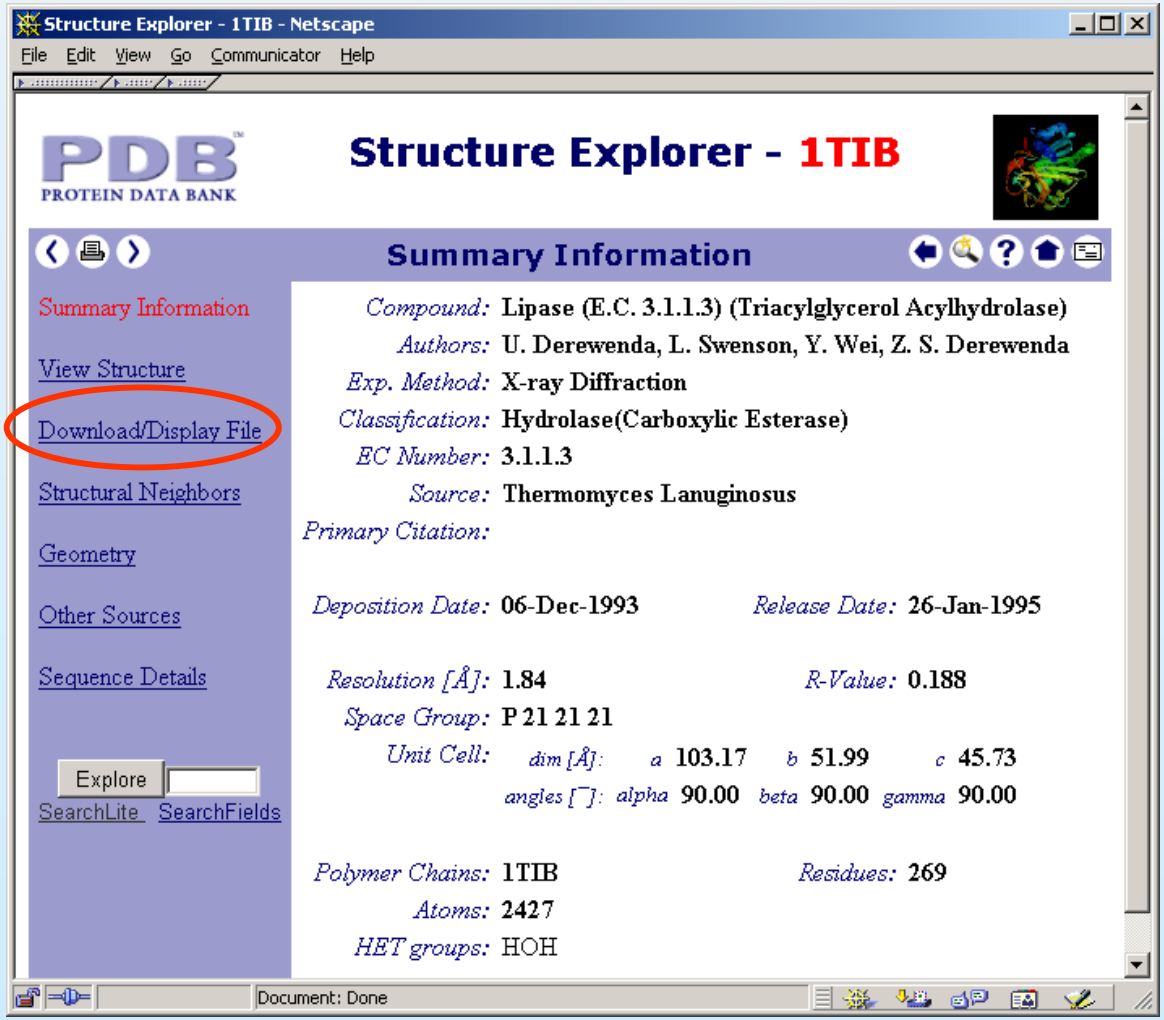
Suchbegriffe

Zugang über Entrez : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide

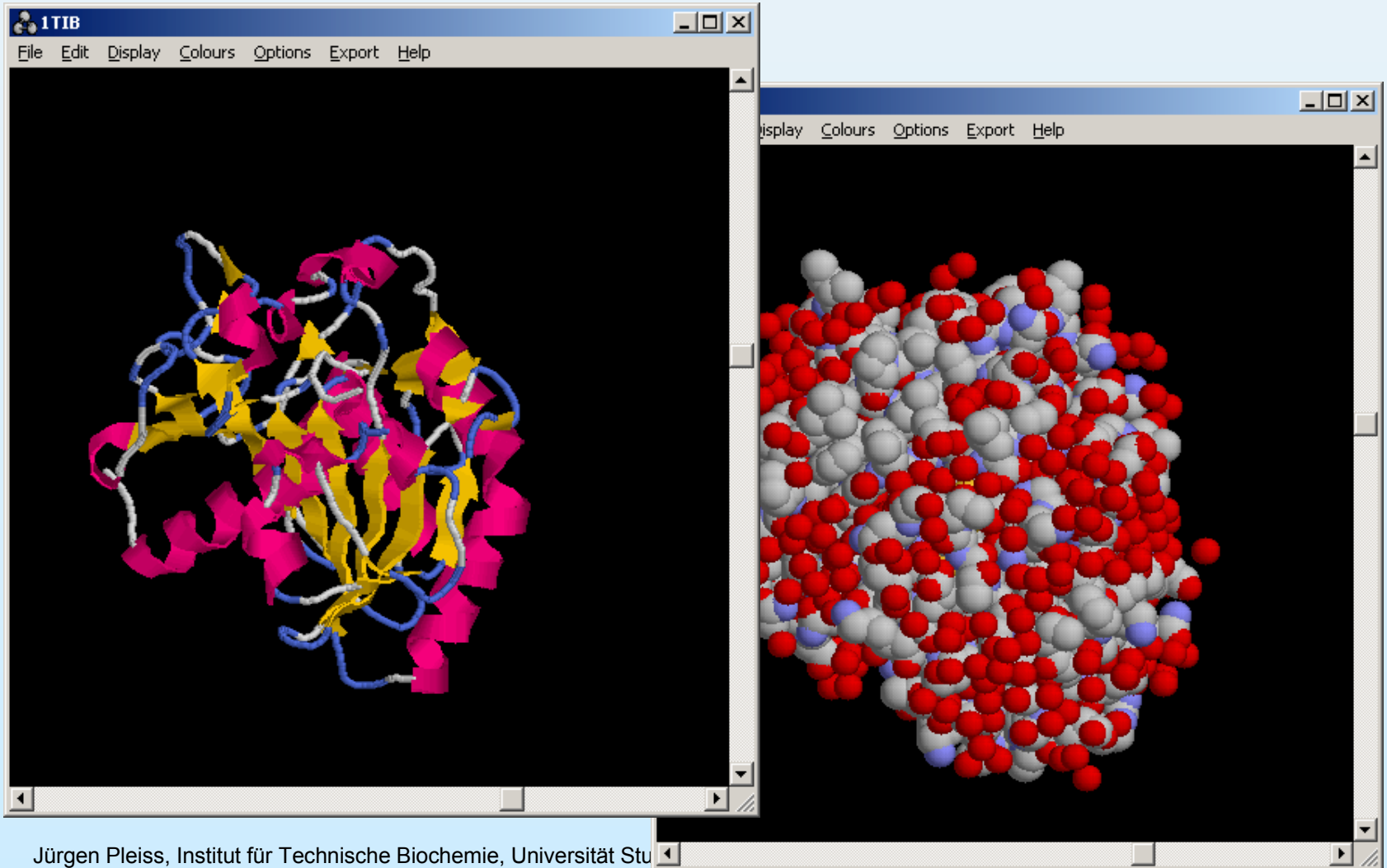
Sequenzdatenbank: GenBank



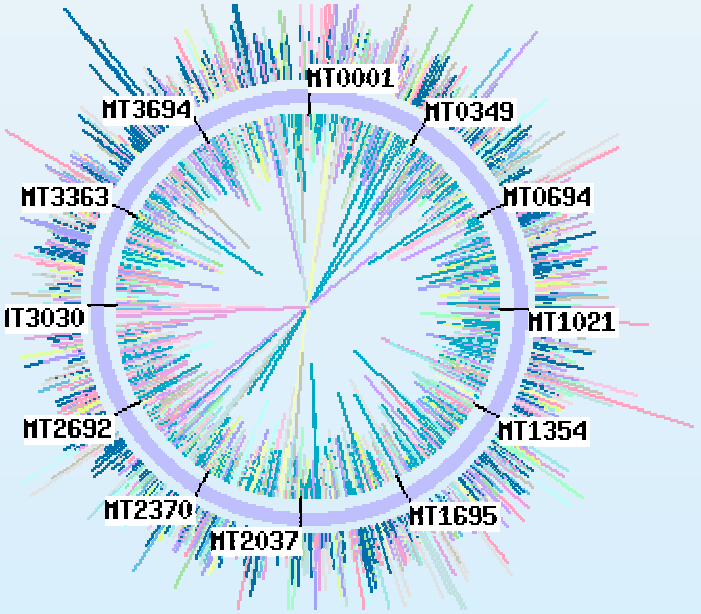
Strukturdatenbank: Protein Data Bank

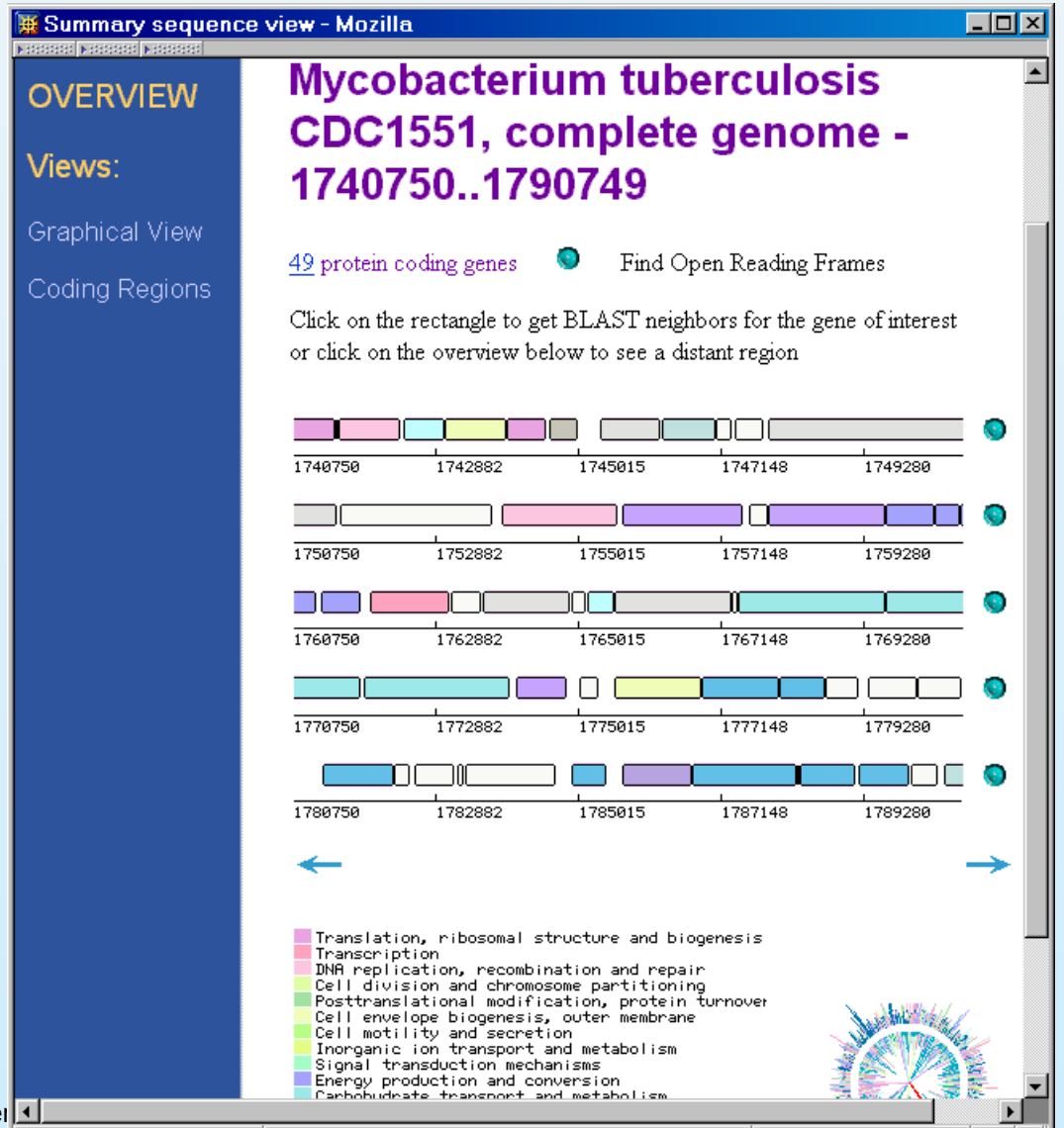


Visualisierung von Strukturen

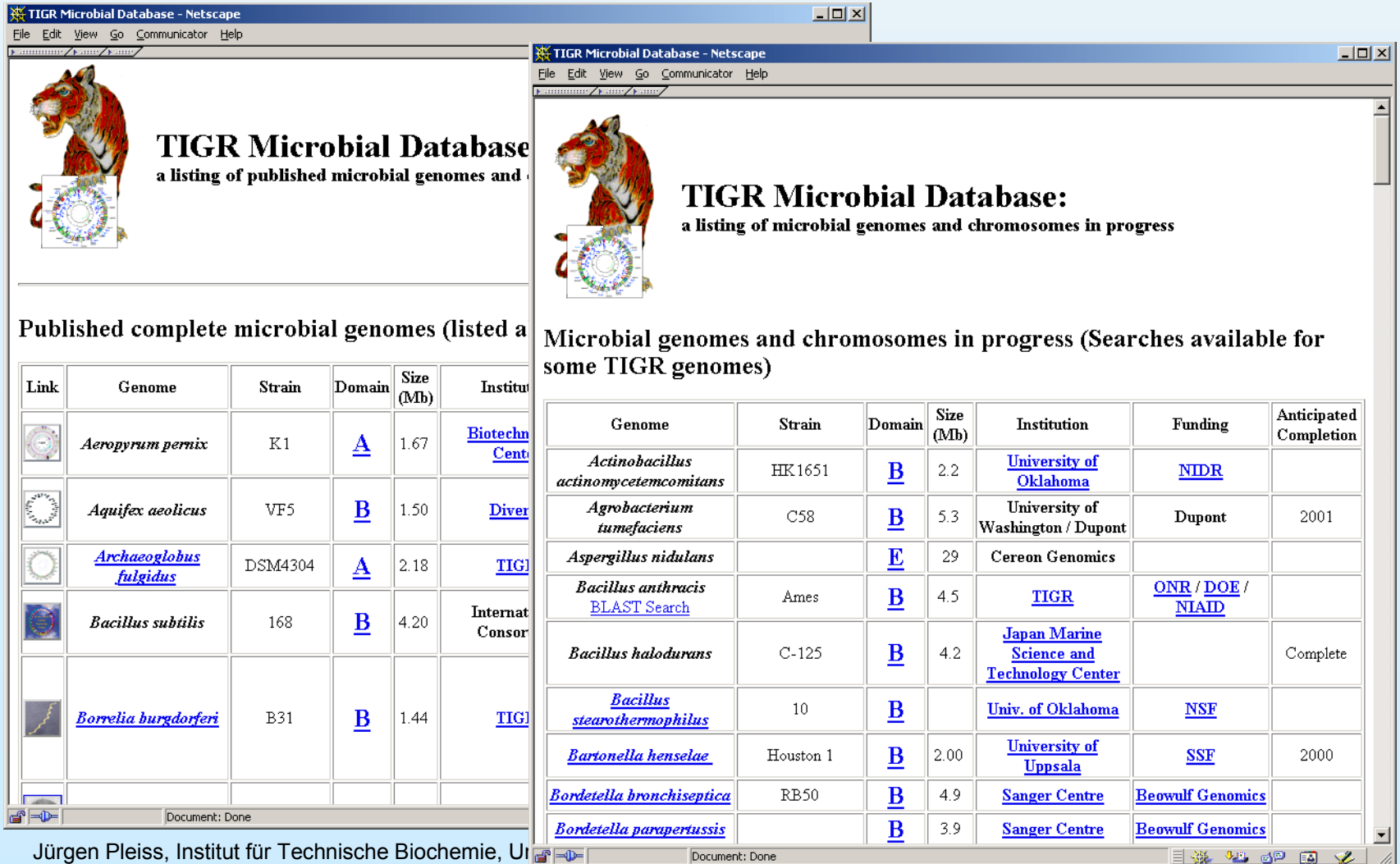


Genomics Datenbanken





Genomics Datenbanken



TIGR Microbial Database - Netscape

File Edit View Go Communicator Help

TIGR Microbial Database

a listing of published microbial genomes and

Published complete microbial genomes (listed a

Link	Genome	Strain	Domain	Size (Mb)	Institut
	Aeropyrum pernix	K1	A	1.67	Biotechn Cent
	Aquifex aeolicus	VF5	B	1.50	Diver
	Archaeoglobus fulgidus	DSM4304	A	2.18	TIGI
	Bacillus subtilis	168	B	4.20	Internat Consor
	Borrelia burgdorferi	B31	B	1.44	TIGI

Document: Done

TIGR Microbial Database - Netscape

File Edit View Go Communicator Help

TIGR Microbial Database:

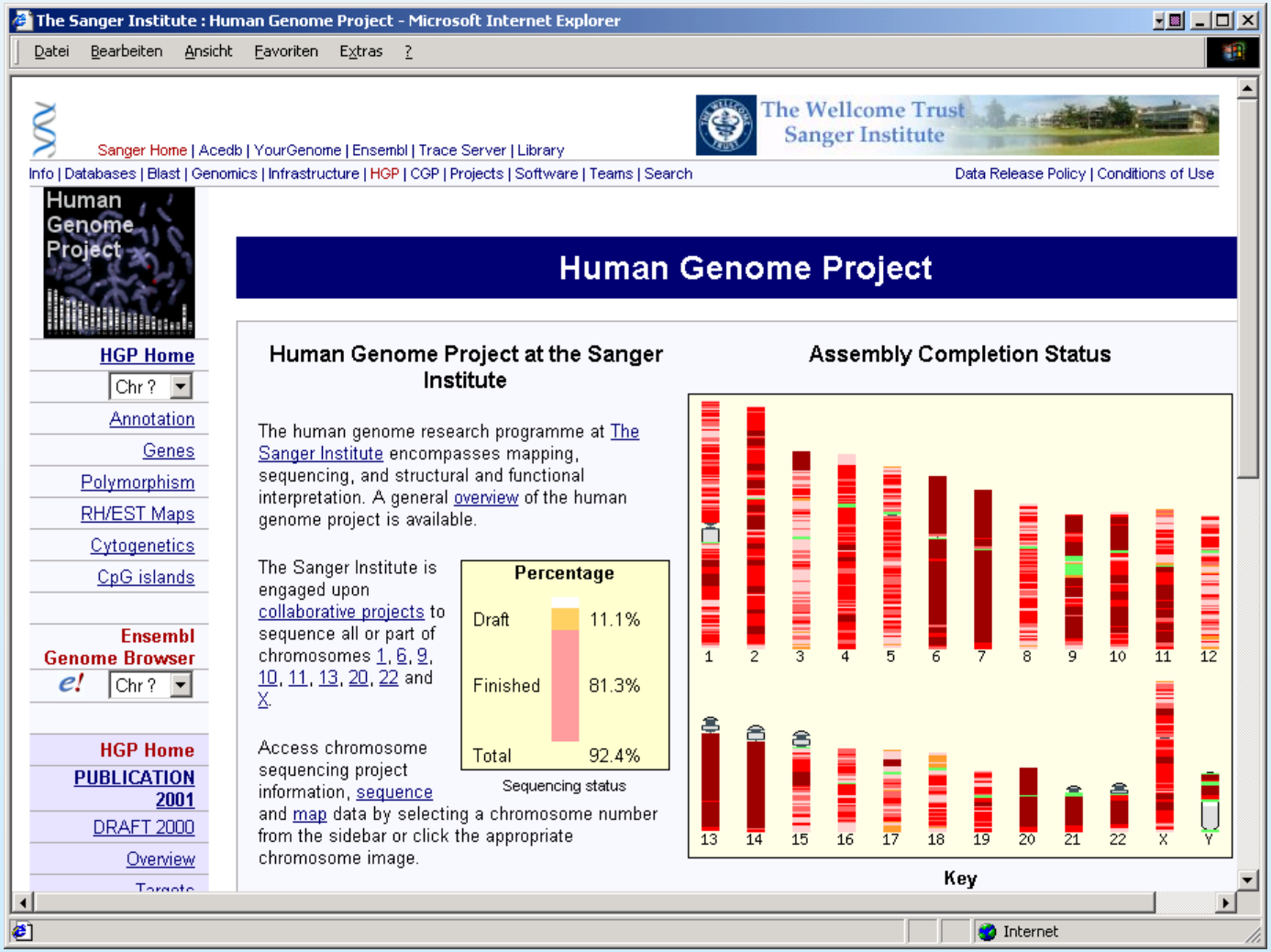
a listing of microbial genomes and chromosomes in progress

Microbial genomes and chromosomes in progress (Searches available for some TIGR genomes)

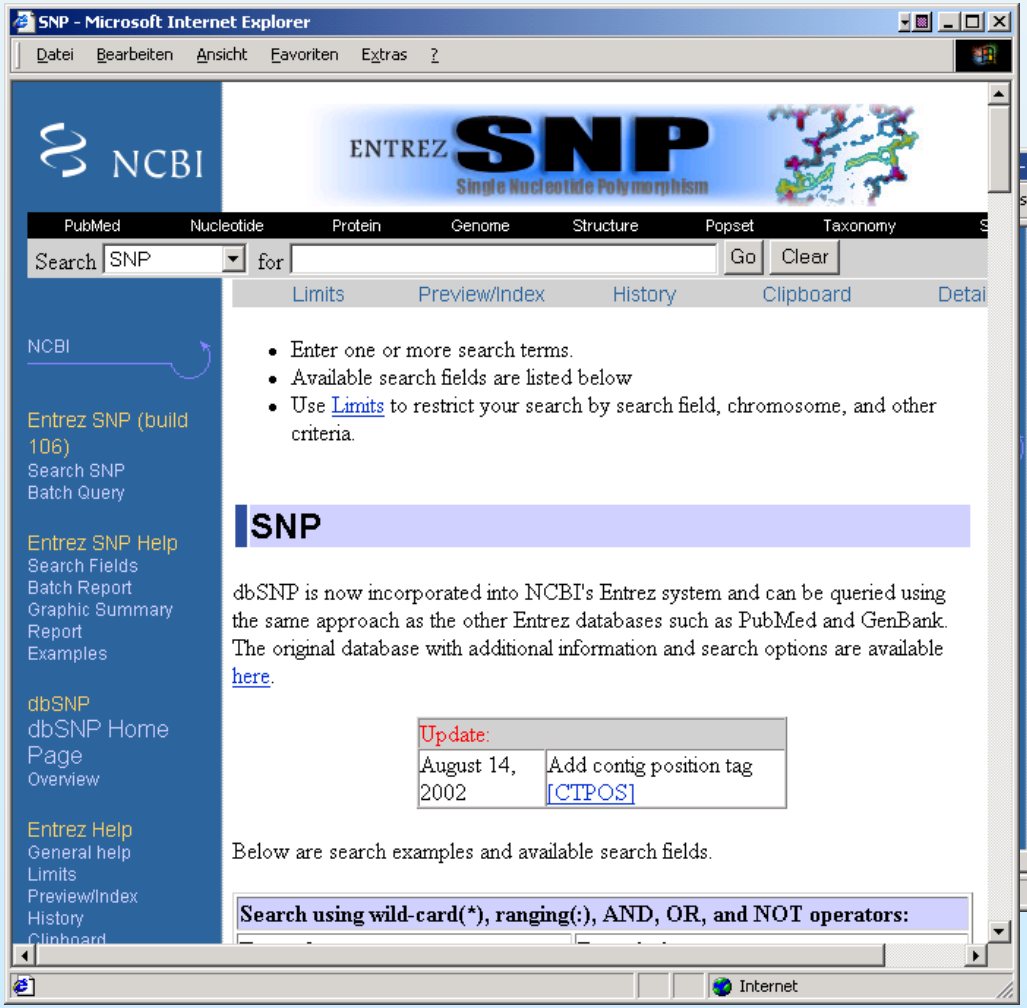
Genome	Strain	Domain	Size (Mb)	Institution	Funding	Anticipated Completion
Actinobacillus actinomycetemcomitans	HK1651	B	2.2	University of Oklahoma	NIDR	
Agrobacterium tumefaciens	C58	B	5.3	University of Washington / Dupont	Dupont	2001
Aspergillus nidulans		E	29	Cereon Genomics		
Bacillus anthracis BLAST Search	Ames	B	4.5	TIGR	ONR / DOE / NIAID	
Bacillus halodurans	C-125	B	4.2	Japan Marine Science and Technology Center		Complete
Bacillus stearothermophilus	10	B		Univ. of Oklahoma	NSF	
Bartonella henselae	Houston 1	B	2.00	University of Uppsala	SSF	2000
Bordetella bronchiseptica	RB50	B	4.9	Sanger Centre	Beowulf Genomics	
Bordetella parapertussis		B	3.9	Sanger Centre	Beowulf Genomics	

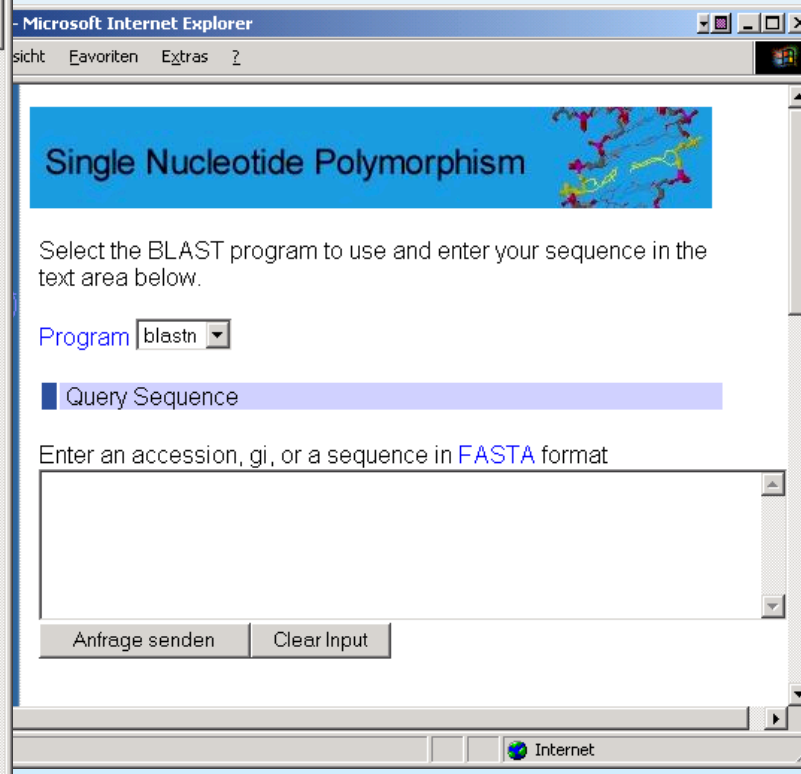
Document: Done

Human Genome Project

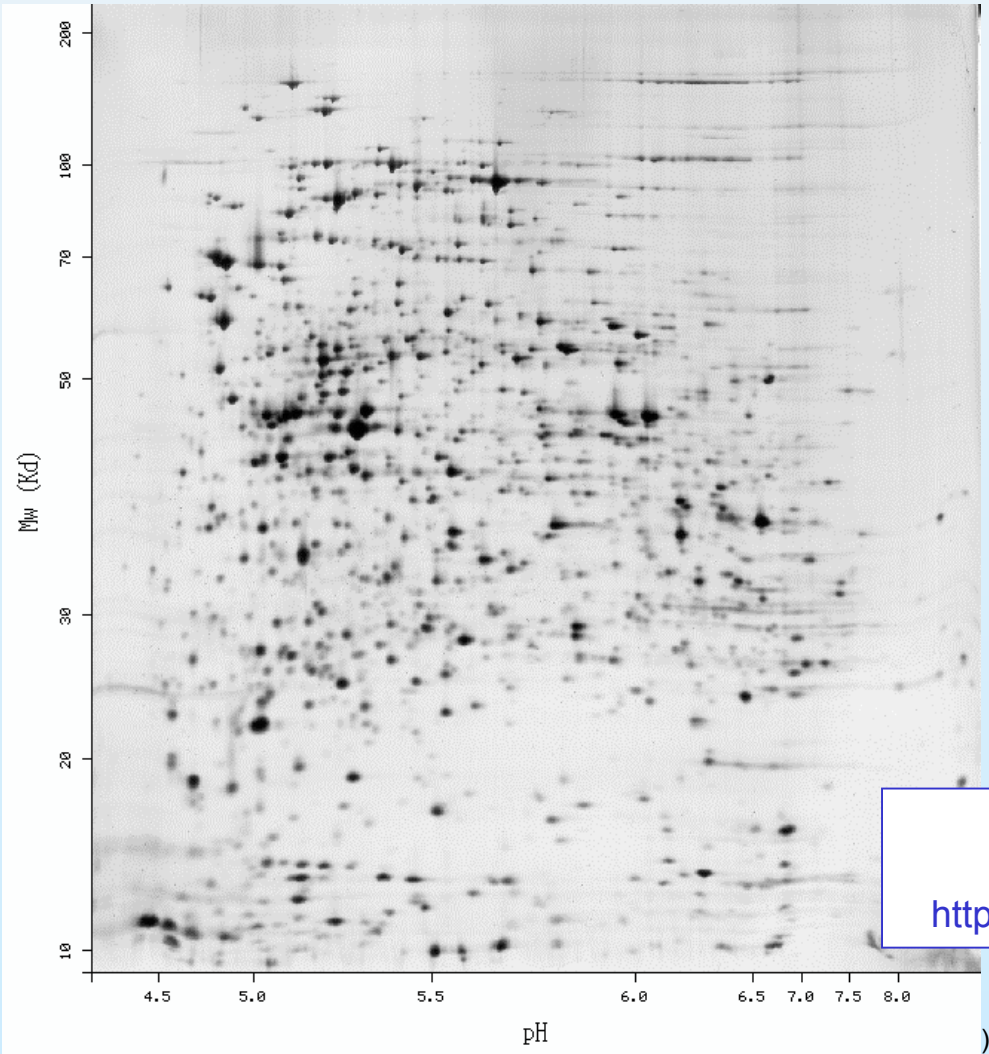


Polymorphismen





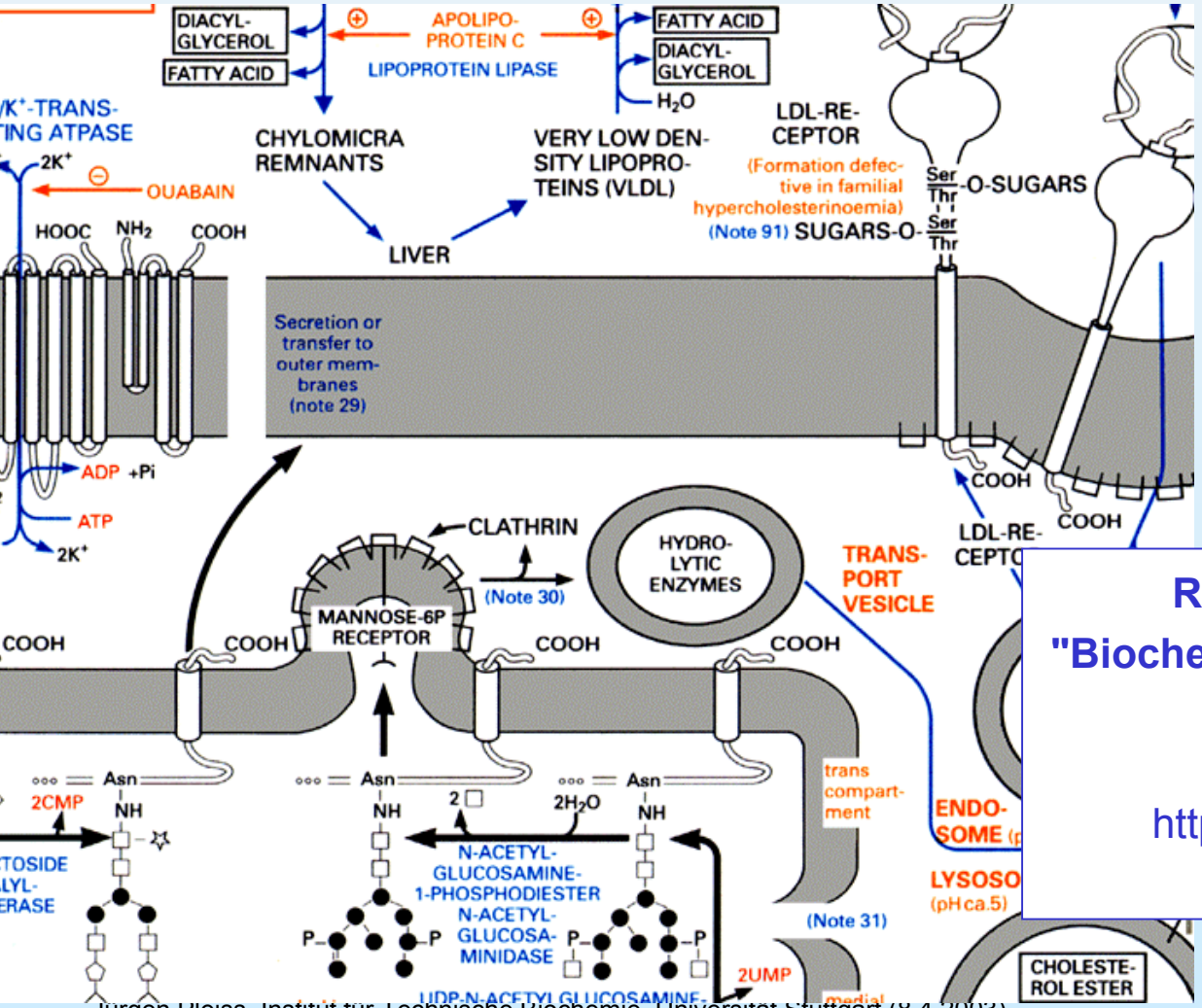
Proteomics



2D Gel von *E.coli* K-12 W3110

http://www.expasy.org/ch2dgifs/ECOLI/ECOLI.gif

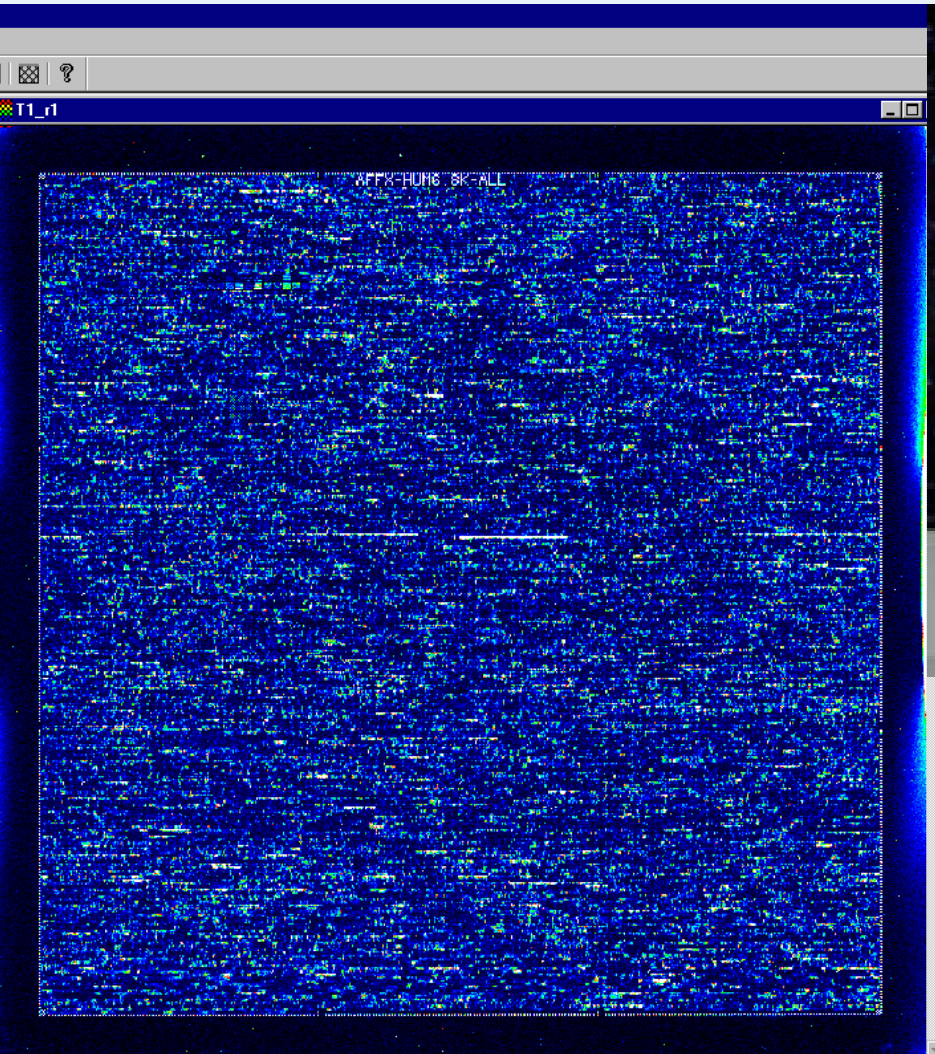
Metabolomics

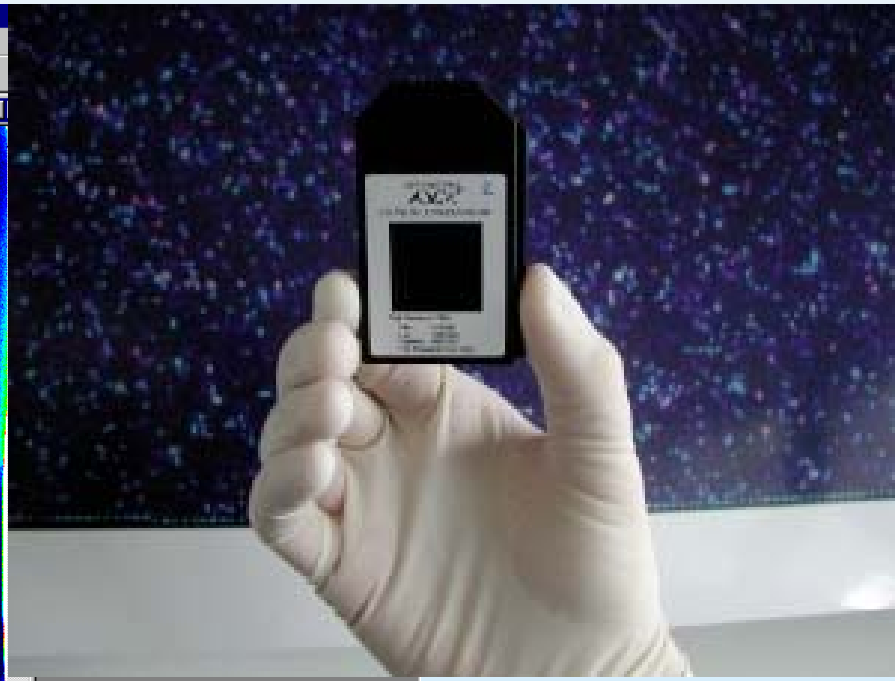


Roche Applied Science
"Biochemical Pathways" wall chart

http://www.expasy.org/cgi-bin/show_image?S3

Transcriptomics





Microarray

(GeneChip expression arrays)

http://www.fgcz.ch/index.php?toc=30

Biologische Daten

Publikationen : Information in Freitext; Stichwörter: Autor, Journal, ...

Datenbanken : wohldefiniertes Datenmodell (z.B. Sequenzdatenbank)

DB-System: ASCII Text oder relationale Datenbank

"discovery driven research": Erzeugung ...

- großer Mengen
- neuer Datentypen
- permanente Änderung der Datenmodelle

Die Grenzen einer Publikation

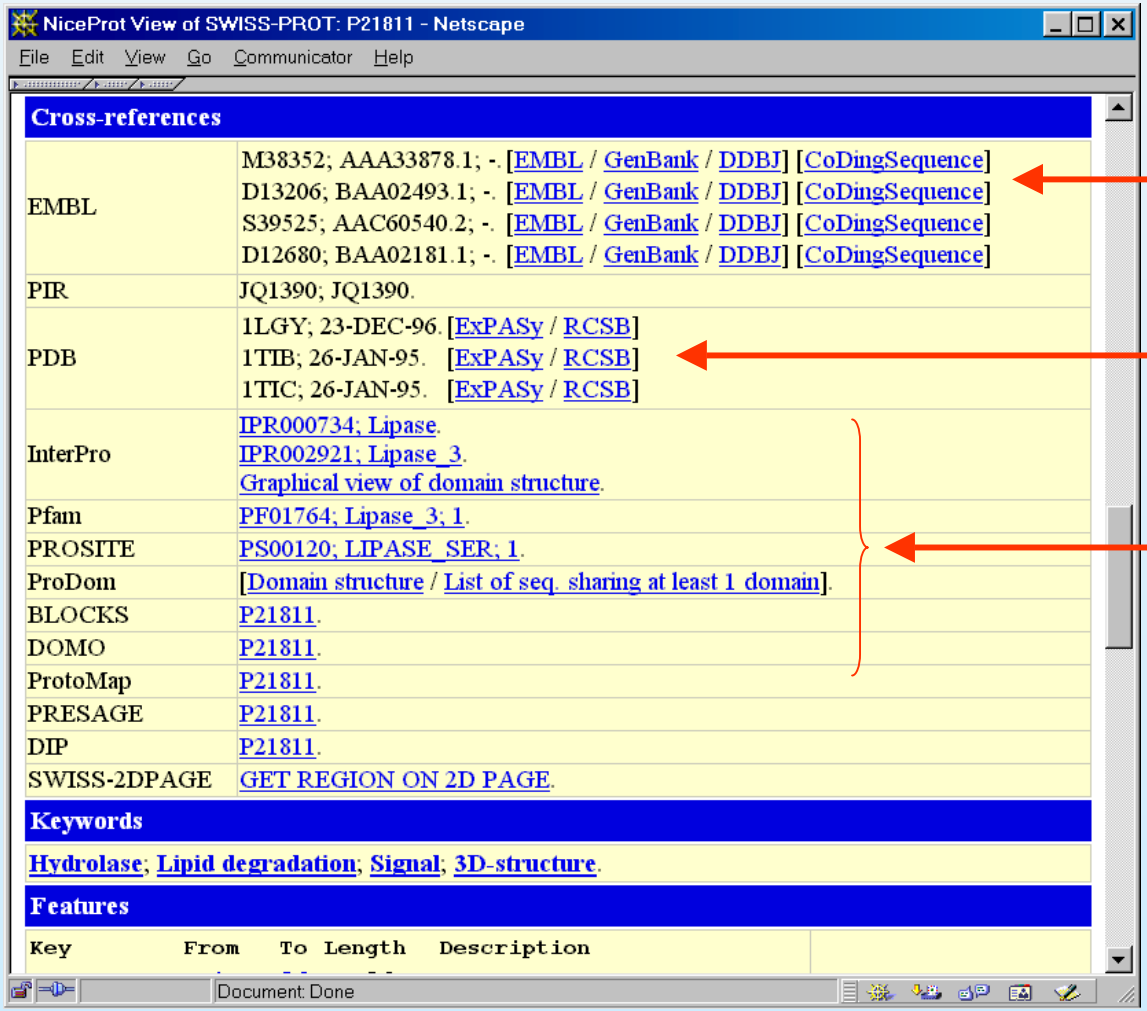
Nature **409**, 942 - 943 (2001)

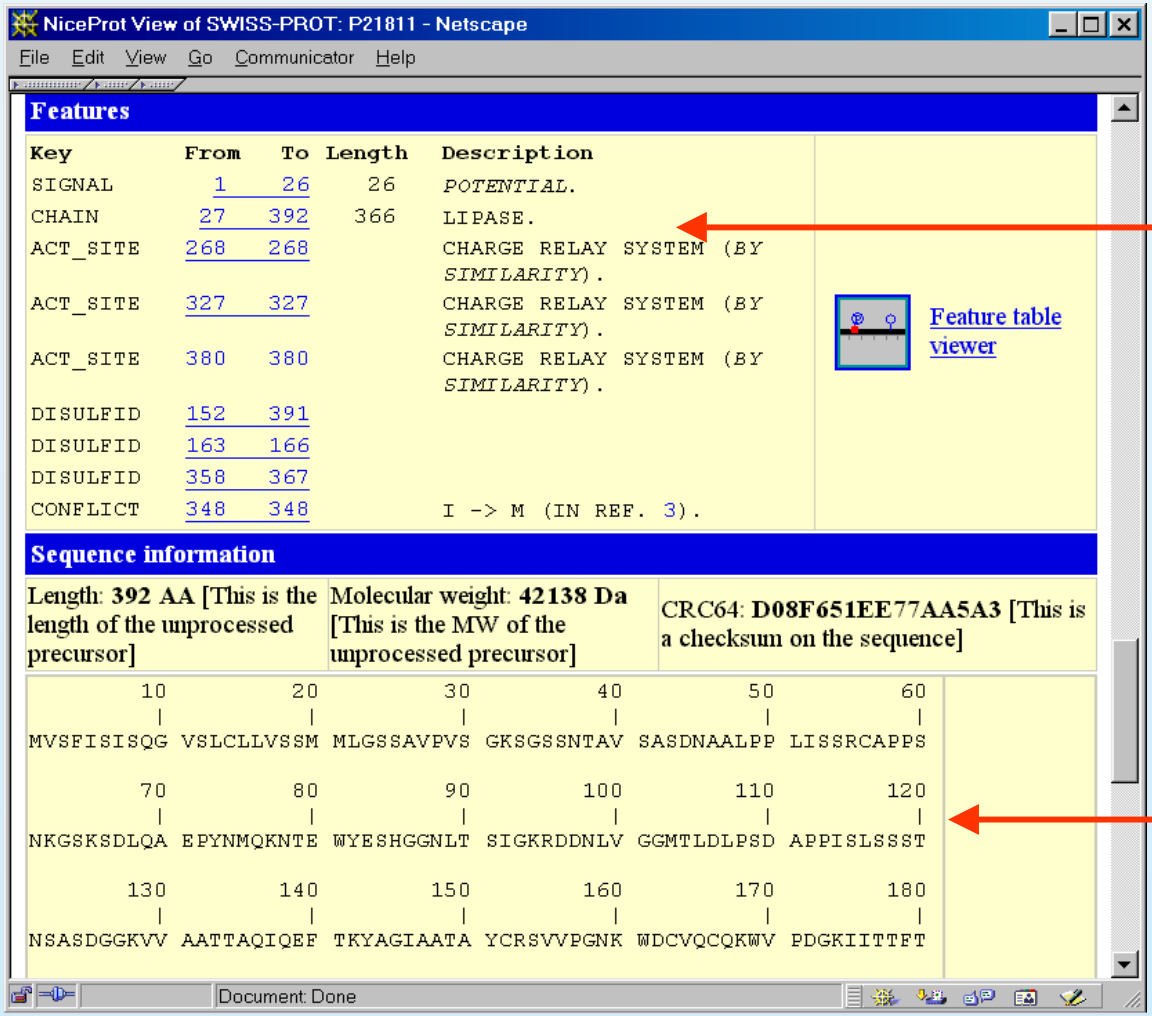
The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X

D. R. BENTLEY*, P. DELOUKAS*, A. DUNHAM*, L. FRENCH*, S. G. GREGORY*, S. J. HUMPHRAY*, A. J. MUNGALL*, M. T. ROSS*, N. P. CARTER*, I. DUNHAM*, C. E. SCOTT*, K. J. ASHCROFT*, A. L. ATKINSON*, K. AUBIN*, D. M. BEARE*, G. BETHEL*, N. BRADY*, J. C. BROOK*, D. C. BURFORD*, W. D. BURRILL*, C. BURROWS*, A. P. BUTLER*, C. CARDER*, J. J. CATANESE†, C. M. CLEE*, S. M. CLEGG*, V. COBLEY*, A. J. COFFEY*, C. G. COLE*, J. E. COLLINS*, J. S. CONQUER*, R. A. COOPER*, K. M. CULLEY*, E. DAWSON*, F. L. DEARDEN*, R. M. DURBIN*, P. J. DE JONG†, P. D. DHAMI*, M. E. EARTHROWL*, C. A. EDWARDS*, R. S. EVANS*, C. J. GILLSON*, J. GHORI*, L. GREEN*, R. GWILLIAM*, K. S. HALLS*, S. HAMMOND*, G. L. HARPER*, R. W. HEATHCOTT*, J. L. HOLDEN*, E. HOLLOWAY*, B. L. HOPKINS*, P. J. HOWARD*, G. R. HOWELL*, E. J. HUCKLE*, J. HUGHES*, P. J. HUNT*, S. E. HUNT*, M. IZMAJLOWICZ*, C. A. JONES*, S. S. JOSEPH*, G. LAIRD*, C. F. LANGFORD*, M. H. LEHVASLAIHO*, M. A. LEVERSHA*, O. T. MCCANN*, L. M. MCDONALD*, J. MCDOWALL*, G. L. MASLEN*, D. MISTRY*, N. K. MOSCHONAS‡, V. NEOCLEOUS§, D. M. PEARSON*, K. J. PHILLIPS*, K. M. PORTER*, S. R. PRATHALINGAM*, Y. H. RAMSEY*, S. A. RANBY*, C. M. RICE*, J. ROGERS*, L. J. ROGERS*, T. SARAFIDOU‡, D. J. SCOTT*, G. J. SHARP*, C. J. SHAW-SMITH*, L. J. SMINK*, C. SODERLUND*, E. C. SOTHERAN*, H. E. STEINGRUBER*, J. E. SULSTON*, A. TAYLOR*, R. G. TAYLOR*, A. A. THORPE*, E. TINSLEY*, G. L. WARRY*, A. WHITTAKER*, P. WHITTAKER*, S. H. WILLIAMS*, T. E. WILMER*, R. WOOSTER* & C. L. WRIGHT*

* The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
‡ Department of Biology, University of Crete and Institute of Molecular Biology and Biotechnology, PO Box 2208, 71409 Heraklion, Crete, Greece
§ Neurogenetic Laboratory, The Cyprus Institute of Neurology and Genetics, 6, International Airport Avenue, PO Box 23462, 1683 Nicosia, Cyprus
† Children's Hospital-BACPAC Resources, 747 52nd Street, Oakland, California 94609, USA

Publikation der Ergebnisse in Datenbanken




NiceProt View of SWISS-PROT: P21811 - Netscape
File Edit View Go Communicator Help
Features
Key From To Length Description
SIGNAL 1 26 26 POTENTIAL.
CHAIN 27 392 366 LIPASE.
ACT_SITE 268 268 CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE 327 327 CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE 380 380 CHARGE RELAY SYSTEM (BY SIMILARITY).
DISULFID 152 391
DISULFID 163 166
DISULFID 358 367
CONFLICT 348 348 I -> M (IN REF. 3).
Feature table viewer
Sequence information
Length: 392 AA [This is the length of the unprocessed precursor]
Molecular weight: 42138 Da [This is the MW of the unprocessed precursor]
CRC64: D08F651EE77AA5A3 [This is a checksum on the sequence]
10 20 30 40 50 60
MVSFISISQG VSLCLLVSSM MLGSSAVPVS GKSGSSNTAV SASDNAALPP LISSRCAPPS
70 80 90 100 110 120
NKGSKSDLQA EPYNMQKNTE WYESHGGNLT SIGKRDDNLV GGMTLDLPSD APPISLSSST
130 140 150 160 170 180
NSASDGGKVV AATTAQIQEF TKYAGIAATA YCRSVVPGNK WDCVQCQKWV PDGKIITTFT
Document: Done
Annotation
Sequenz

Anwendungen

- Diagnostik
- Pharmacogenomics (Korrelation Genotyp-Phänotyp)
- Protein Engineering
- Metabolic Engineering

Welche Daten werden erzeugt und ausgetauscht?

- Experimentelle Rohdaten oder biologisch interpretierte Daten

  Informationen über Versuchsbedingungen (LIMS)

  Ergebnisse (ASCII, Bildmaterial)
- Textbasierte Interpretation
- Programme zur Auswertung

- Austausch: World Wide Web
- Problem: Standardisierung (Format; Datenmodell)

Datenintegration

Datenbanken: primäre, sekundäre Daten

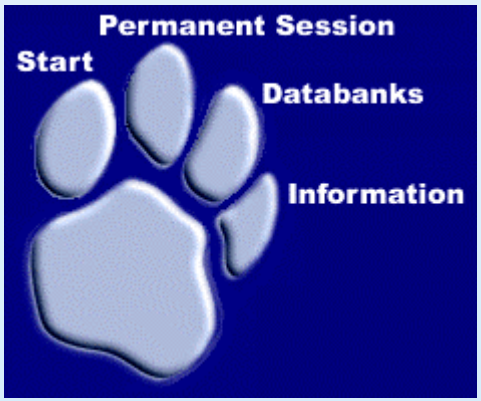
Links zu Einträgen in anderen Datenbanken (Bsp: SwissProt)

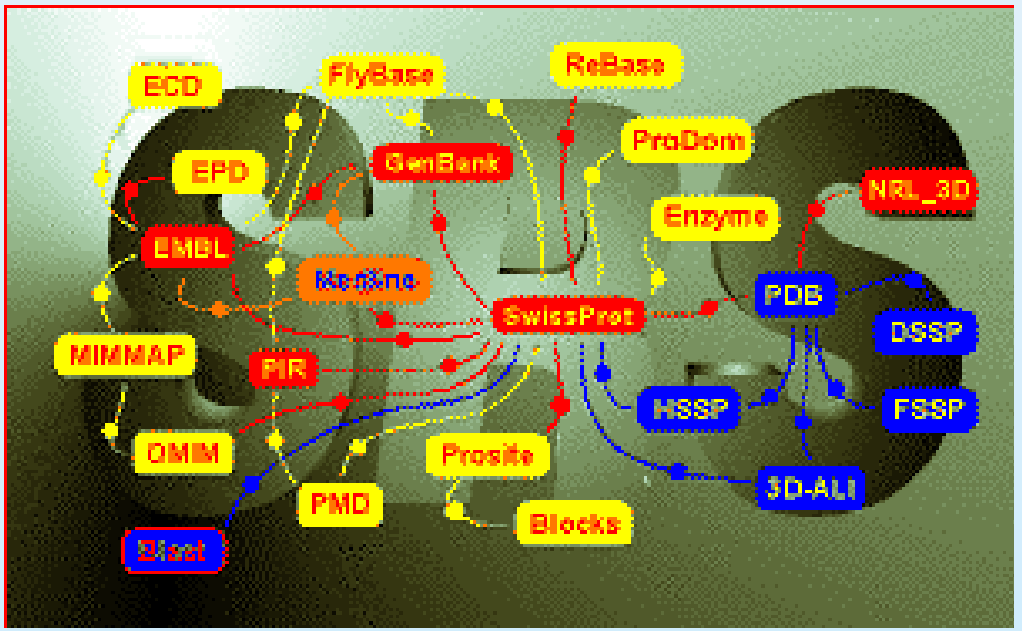
Integration von Datenbanken

Data Warehouse | Federated Database System | BioXML

Datenintegration : das Sequence Retrieval System

http://srs.ebi.ac.uk/ → Integrierter Zugang zu 140 Datenbanken





Verwendung der Daten

Forschung

- Verknüpfung mit eigenen Daten
- Analyse von (fremden) Rohdaten unter unterschiedlichen Aspekten

  → Theoriebildung

Lehre (Vorlesung, Praktika)

- zusätzlich zu klassischem Lehrmaterial (Lehrbücher, Skripte): interaktives und über das Internet zugängliches Lehrmaterial
- Möglichkeit, im Rahmen von Praktika auf Originaldaten(-banken) zuzugreifen (nicht nur reduzierte Versuchsaufbauten)

Wie wird publiziert?

Publikation von Forschungsergebnissen

- Publikation in "peer-reviewed journals" (impact factor!); zusätzlich:
- supplemental material (Daten, Programme) auf server des Verlags
- zusätzliche Daten auf Anfrage bei den Autoren
- Verweis auf eigene homepage (Problem: keine Langzeitsicherung)
- Einträge in externen, "offiziellen" Datenbanken (z.B. Sequenz, Struktur)
- Betrieb einer eigenen, langfristigen Datenbank (commitment)

Diplom- und Doktorarbeiten: ausschließlich Druckversion oder E-Print

Wie werden Daten gesucht?

Literatur-Recherche (Medline, CA)

nach Stichwort oder Autor

wer hat eine gegebene Arbeit zitiert?

Volltext-Zugang: online journal

lokale Bibliothek

SUBITO

Suche in online-Angeboten:

Verweise in Publikationen oder home pages

Portale von Fachgruppen oder Arbeitsgruppen

allgemeine Suchmaschinen (Google)

Trends

- Literatur muß online zugänglich sein
- Zunahme der Relevanz von Rohdaten
- home pages der Forschergruppen zusätzlich zu "offiziellen" Datenbanken

Problem: Datenbereitstellung über langen Zeitraum (Volatilität der Daten)

Zugänglichkeit der Daten / Datensicherheit